Cognitive AI-Based Emotional State Monitoring for Counselling Sessions

M. Monika[1], M. Abinaya[2], T. Bavana Sri[3], S. Pooja[4]

Department of Information Technology, Syed Ammal Arts and Science College, Ramanathapuram, India

[1]monikapriya091@gmail.com | [2]abinayamurugan078@gmail.com | [3]bavanasri820@gmail.com | [4]poojaselvam54@gmail.com

Abstract— Mental health counselling requires continuous understanding of an individual's emotional and behavioural changes. This paper presents a Cognitive AI-based Emotional State Monitoring System that is designed to analyse observable emotional cues from facial expressions, speech, and conversational text. The framework considers emotions such as happiness, sadness, anger, fear, surprise, and neutrality and tracks their variation during counselling sessions. A multimodal pipeline combines modality-specific predictions through late fusion and applies temporal analysis to identify emotional transitions. The resulting indicators are presented through a counsellor dashboard as supplementary decision-support information. The system is explicitly positioned as a support tool rather than a clinical diagnostic system; final interpretation remains with qualified mental health professionals. The current manuscript describes the proposed architecture and evaluation framework, while the numerical comparison included in the source document is retained as illustrative benchmark data and should be replaced by measured experimental results before publication.

Keywords— Cognitive AI, Emotion Recognition, Facial Emotion Recognition, Deep Learning, Mental Health, Multimodal Emotion Analysis, Explainable AI.

I. INTRODUCTION

Mental health counselling is a complex and continuous process that requires an effective understanding of an individual's emotional and behavioural changes. During counselling sessions, individuals may communicate through facial expressions, speech, and written or verbal language. These cues are not always explicit or straightforward to interpret through conventional observation alone. Artificial Intelligence (AI) therefore provides an opportunity to support counsellors by analysing observable emotional patterns in a systematic and consistent manner.

Recent developments in Cognitive AI, Deep Learning, Computer Vision, and Emotion Recognition have enabled intelligent systems to analyse emotional cues from multiple sources. Facial Emotion Recognition (FER) is commonly used to classify observable facial-expression categories such as happiness, sadness, anger, fear, surprise, and neutrality [1]. Deep learning models can extract discriminative facial features and learn complex patterns, making them suitable for real-time emotion-monitoring applications.

The proposed Cognitive AI-Based Emotional State Monitoring System focuses on analysing emotional variation throughout a counselling session and presenting structured indicators through a counsellor dashboard. Facial, speech, and textual information can be processed independently and subsequently integrated to provide a multimodal representation. Temporal analysis is included because an emotional state may change during a session rather than remain constant.

Because emotions are subjective and context-dependent, automated emotion recognition should not be treated as a direct measure of a person's clinical or psychological condition. The proposed system is therefore intended as decision support, with professional interpretation and counselling decisions remaining under the responsibility of qualified mental health professionals. This human-centred positioning is essential for responsible use of AI in counselling environments.

II. RELATED WORK

2.1 Facial Emotion Recognition in Counselling

Previous research has investigated facial emotion recognition for identifying observable emotional responses during psychotherapy and counselling-related settings. Machine learning and deep learning models have been applied to facial expressions to classify categories such as happiness, sadness, anger, fear, surprise, and neutrality [2]. These approaches demonstrate the feasibility of automated facial analysis; however, facial expressions alone may not provide sufficient context for interpreting an individual's emotional experience.

2.2 AI-Based Mental Health Monitoring

Artificial Intelligence has increasingly been applied to mental-health monitoring through behavioural and emotional-data analysis. Such systems can identify patterns that may assist professionals in understanding changes in observable behaviour. However, AI-based monitoring should remain a supportive capability and should not replace professional counselling or clinical judgement.

2.3 Multimodal Emotion Recognition

Multimodal emotion recognition combines information from facial expressions, speech, and textual or linguistic cues. Integrating multiple modalities can provide a broader representation because verbal and non-verbal signals may complement one another. Multimodal machine learning is therefore relevant to applications in which emotional information is distributed across heterogeneous data sources [6].

2.4 Speech- and Text-Based Emotional Analysis

Speech analysis can capture acoustic characteristics such as pitch, energy, spectral properties, and temporal speaking patterns. Natural Language Processing (NLP) can extract contextual and semantic information from spoken or written language. Combining speech and textual information with facial cues may reveal transitions that are not apparent from a single modality.

2.5 Temporal Emotional-State Analysis

Emotional states may change throughout a counselling session. Temporal modelling can therefore be used to compare successive predictions and identify transitions or persistent patterns. Continuous analysis can provide counsellors with a chronological view of observable emotional indicators rather than isolated predictions.

2.6 Cognitive AI for Counselling Support

Cognitive AI extends conventional emotion recognition by incorporating contextual interpretation, pattern analysis, and decision-support mechanisms. In a counselling environment, such a framework can combine multiple behavioural cues and present structured indicators to professionals. The objective is support for professional awareness rather than autonomous diagnosis.

2.7 Research Gap

The source article identifies several gaps in existing approaches: many systems focus on isolated emotional predictions rather than continuous monitoring of emotional transitions; comparatively limited attention is given to counsellor-oriented visualisation, contextual interpretation, explainability, and privacy-aware implementation. These gaps motivate the proposed multimodal and temporal framework.

2.8 Proposed Research Direction

The proposed system addresses these gaps by integrating multimodal emotional cues and continuously analysing observable emotional changes during counselling sessions. The output is organised through a counsellor dashboard to support professional observation and decision-making while maintaining professional judgement as the primary basis for assessment.

III. PROPOSED COGNITIVE AI-BASED EMOTIONAL STATE MONITORING

The proposed framework is a Cognitive AI-based system designed to monitor observable emotional cues during counselling sessions. It processes facial expressions, speech characteristics, and conversational content, integrates the resulting predictions, and applies temporal analysis to identify changes over time. The final indicators are presented through a counsellor dashboard as supplementary decision-support information. The framework does not provide autonomous clinical diagnosis.

3.1 System Architecture and Workflow

The architecture consists of data acquisition, preprocessing, modality-specific feature extraction, emotion recognition, multimodal fusion, temporal analysis, and counsellor-support layers. Real-time facial video, speech signals, and conversational text are processed through their respective analytical modules. The extracted visual, acoustic, and linguistic features are then integrated to obtain a unified representation.

The workflow begins with real-time data acquisition, followed by modality-specific preprocessing and feature extraction. Emotion recognition models generate modality-level predictions, which are combined using late fusion. A temporal analysis module compares consecutive emotional states and identifies transitions or persistent patterns. The resulting indicators are displayed on the counsellor dashboard for professional interpretation.

Real-Time Input → Preprocessing → Feature Extraction → Emotion Recognition → Multimodal Fusion → Temporal Analysis → Emotional-State Estimation → Counsellor Dashboard → Professional Interpretation

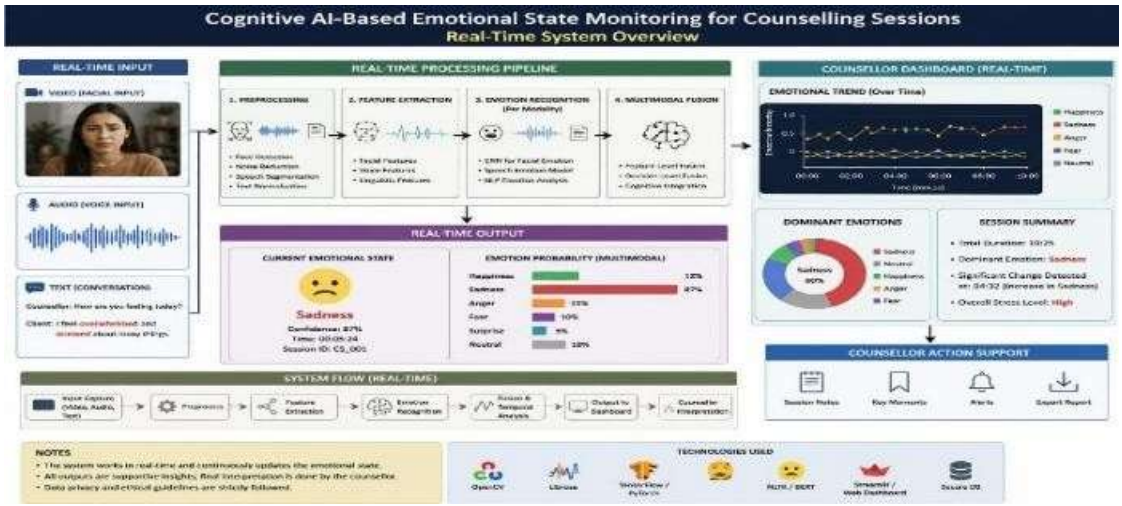


Fig. 1. Illustrative counsellor dashboard for monitoring observable emotional-state indicators.

3.2 Proposed Real-Time Multimodal Emotional-State Monitoring Algorithm

Input: Real-time facial video, speech signal, and conversational text.

Output: Real-time observable emotional-state estimate, confidence score, and temporal emotional trend.

1) System initialization: Initialize the camera, microphone, speech-processing and NLP modules, modality-specific models, temporal model, and counsellor dashboard.

2) Real-time data acquisition: Continuously capture facial video frames, speech signals, and conversational text from the counselling session.

3) Input synchronization: Synchronize multimodal inputs using timestamps to maintain temporal correspondence among facial, speech, and textual information.

4) Face detection: Detect and isolate the participant's facial region from each incoming video frame.

5) Video preprocessing: Resize frames and perform normalization, noise reduction, and illumination adjustment.

6) Speech preprocessing: Segment the audio stream, reduce background noise, and convert the signal into suitable acoustic representations.

7) Text preprocessing: Convert available speech or conversation into structured text and perform tokenization and normalization.

8) Facial feature extraction: Use a deep convolutional model (specified in the source workflow as ResNet-50) to extract spatial facial features.

9) Speech feature extraction: Extract acoustic characteristics such as pitch, energy, spectral features, and temporal variations.

10) Speech emotion recognition: Apply a CNN-LSTM model in which convolutional layers capture local acoustic patterns and LSTM layers model temporal dependencies [3], [4].

11) Linguistic feature extraction: Use BERT to capture contextual and semantic information from the counselling conversation [5].

12) Modality-level prediction: Generate independent emotion predictions from the facial, speech, and text models.

13) Confidence estimation: Calculate a confidence score for each modality-level prediction to represent model confidence.

14) Multimodal fusion: Combine the facial, speech, and textual predictions using a late-fusion mechanism.

15) Temporal emotion analysis: Apply an LSTM-based temporal module to compare consecutive emotional states and identify transitions.

16) Real-time state estimation: Determine the current dominant observable emotional state from the fused multimodal representation.

17) Continuous update: Update the estimated state whenever new synchronized multimodal input becomes available.

18) Significant-change detection: Identify substantial changes in the observed emotional pattern and mark them as analytical indicators.

19) Dashboard visualisation: Display the current state, confidence score, and temporal trend on the counsellor dashboard.

20) Session-level analysis: Maintain a chronological record of detected states and generate an overall session-level trend.

21) Counsellor decision support: Provide the generated indicators as supplementary information for professional observation and interpretation.

22) Privacy and security: Apply controlled access, secure data handling, and appropriate privacy measures for sensitive counselling information.

23) Session termination: Stop real-time monitoring at the end of the session and generate a structured session-level summary.

IV. PERFORMANCE EVALUATION AND COMPARATIVE ANALYSIS

The proposed framework can be evaluated using three principal modalities: facial expressions, speech, and counselling text. The source document specifies Accuracy, Precision, Recall, F1-Score, and the Confusion Matrix as evaluation measures. Because the source article does not provide a dataset description, trained-model files, or measured experimental logs, the percentage values retained below are explicitly labelled as illustrative values rather than verified experimental results.

4.1 Models Used

Modality	Model / Approach	Purpose
Facial emotion	ResNet-50 / CNN	Extract facial features and classify observable facial-expression categories.
Speech emotion	CNN-LSTM	Model acoustic features and their

		temporal dependencies.
Text emotion	BERT	Capture contextual and semantic information from counselling text.
Multimodal emotion	Late-fusion ensemble	Combine facial, speech, and text predictions.

4.2 Training and Testing

The source article specifies an 80% training and 20% testing split. The corresponding workflow is:

Dataset → Preprocessing → Feature Extraction → Model Training → Validation
Test Data → Trained Model → Emotion Prediction → Performance Evaluation

4.3 Performance Metrics

- Accuracy: proportion of correctly classified instances among all evaluated instances.
- Precision: proportion of predicted instances of an emotion that are correctly classified.
- Recall: proportion of actual instances of an emotion that are correctly identified.
- F1-score: harmonic mean of precision and recall.
- Confusion matrix: class-wise summary of correct and incorrect predictions.

4.4 Illustrative Comparative Values

The following values were present in the source document and were described there as sample values. They must not be reported as measured experimental results. They are retained only to preserve the source content and to indicate the intended comparative-evaluation format.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	84%	83%	82%	82%
SVM	88%	87%	87%	87%
Random Forest	86%	85%	84%	84%
CNN	92%	91%	91%	91%
CNN-LSTM	94%	93%	93%	93%
Multimodal Fusion	96%	95%	95%	95%

Note: Illustrative values from the source document; replace with measured results obtained on the final dataset before submission.

4.5 Comparative Interpretation

For the final experimental study, model comparison should be based on the same dataset partitioning, preprocessing protocol, emotion classes, and evaluation procedure. The multimodal model should be assessed against each single-modality baseline using identical test data and clearly reported confidence intervals or repeated-run statistics where applicable. A final model-selection statement should be made only after measured results are available.

V. REAL-TIME IMPLEMENTATION AND COUNSELLOR SUPPORT

The trained models can subsequently be integrated into a real-time counselling support system. The intended operational workflow is:

Camera + Microphone + Counselling Text
↓
Facial Emotion Detection + Speech Emotion Detection + Text Emotion Analysis
↓
Multimodal Fusion
↓
Final Observable Emotional State
↓
Counsellor Dashboard

The dashboard may display the current observable emotional state, model confidence, emotion variation over the session, emotion trends, detected changes, and a session-level summary. These outputs should be presented as supplementary information, with the counsellor retaining responsibility for interpretation.

5.1 Future Development

- Real-time implementation: integrate camera, microphone, and text inputs for synchronized emotion monitoring.
- Advanced deep learning: evaluate CNN-LSTM, Transformer, and BERT-based architectures under a common experimental protocol.
- Multimodal fusion: investigate alternative fusion mechanisms and quantify their contribution over single-modality baselines.
- Emotion trend analysis: generate a session-level timeline of observable emotional-state changes.
- Counsellor dashboard: provide interpretable displays of detected states, confidence scores, trends, and session summaries.
- Explainable AI: provide model-level or feature-level explanations for predicted emotional categories where technically appropriate.
- Larger and diverse datasets: evaluate generalisation across participants, environments, demographics, and recording conditions.
- Privacy and security: apply secure storage, access control, anonymisation, and privacy-preserving processing for sensitive data.
- Cloud/IoT integration: investigate secure cloud services for session storage and remote access where appropriate.
- Clinical validation: evaluate practical usefulness and reliability with qualified counselling professionals before real-world deployment.

VI. CONCLUSION

This paper presents a Cognitive AI-Based Emotional State Monitoring framework for counselling sessions that combines facial expression analysis, speech emotion recognition, and text-based emotion analysis. The proposed workflow integrates modality-specific predictions through late fusion and applies temporal analysis to monitor changes over the course of a session. A counsellor dashboard is used to organise the resulting indicators for professional observation.

The framework is intended as an AI-based decision-support mechanism rather than an autonomous diagnostic system. The current source document defines the architecture, algorithm, evaluation measures, and an illustrative comparative table, but it does not provide sufficient experimental evidence to validate the reported percentage values. Therefore, actual dataset-based training, testing, ablation analysis, confusion matrices, and real-time validation are required before the performance claims can be treated as experimental findings.

Future work should focus on validated multimodal datasets, temporal evaluation, explainability, privacy-preserving implementation, and evaluation with qualified counselling professionals. These steps can establish the practical usefulness and limitations of the proposed framework while preserving professional oversight.

REFERENCES

[1] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[2] P. Ekman, "An argument for basic emotions," Cognition and Emotion, 1992.

[3] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," Language Resources and Evaluation, 2008.

[4] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," PLOS ONE, 2018.

[5] J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019.

[6] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.